基于稀疏自编码特征聚类算法的图像窜改检测 *

王梦思，霍宏涛，罗霄阳

(中国人民公安大学 信息技术与网络安全学院，北京 100038)

摘　要：同图复制窜改是图像窜改较为常见的一类，基于块匹配检测方法往往存在准确率低、时间复杂度高等问题，为提高准确率并大幅度降低时间复杂度，应用深度学习特征和聚类算法进行检测。首先用稀疏自编码器训练大量样本集找出同图复制图像的内部规律并得到降维的隐藏层权值矩阵，通过权值矩阵获得检测图像的隐藏层特征，即定义的稀疏自编码特征；用 K-means 算法一次聚类自编码特征去除图像平滑区域，二次聚类纹理特征获得检测结果，若检测结果中含有少量异常块，通过欧氏距离判断和 RANSAC（random sample consensus）算法将异常块去除，从而实现窜改区域的检测。实验结果表明，该算法与其他算法比较综合准确率提升 14.3%，时间效率提升 72%。将深度学习特征与聚类算法结合使用，使得同图复制窜改在时间效率和准确率上皆有所提升。

关键词：稀疏自编码；K-means 聚类算法；同图复制；块匹配

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2017.07.0681

Image forgery detection based on sparse autoencoder feature and clustering algorithm

Wang Mengsi, Huo Hongtao, Luo Xiaoyang

(*Institute of Information Technology & Network Security*, *People's Public Security University of China*, *Beijing* 100038, *China*)

Abstract: Copy-move forgery is a common type of image forgery. Block matching detection often has the problems of low accuracy and high time complexity. In order to improve the accuracy and significantly reduce the time complexity, this paper used deep learning characteristics and clustering algorithm for detecting. Firstly, it used the sparse autoencoder to find out the internal laws of the images and train the weight matrix of the hidden layer which obtained by a large number of sample sets. It obtained the hidden layer feature of the detection image by the weight matrix, that is, the sparse autoencoder feature. Secondly, it used the K-means algorithm to cluster the autoencoder features at the first time to remove the image smoothing region and to cluster the texture features to obtain the detection results. It used the Euclidean distance judgment and RANSAC（random sample consensus）algorithm to remove the abnormal blocks, in order to achieve tampering area detection. Experimental results show that the proposed algorithm can improve the accuracy by about 14.3% compared with other algorithms, and the time efficiency is improved by72%. The combination of the depth learning feature and the clustering algorithm makes the tampering of the copy-move forgery improved in both time efficiency and accuracy.

Key Words: sparse autoencoder; K-means clustering algorithm; copy-move forgery; block matching algorithm

0　引言

随着信息技术的发展，图像编辑软件的广泛使用，图像窜改已变得平常和普遍，其中同图复制窜改是最常见的一类。将图中一个目标体复制粘贴到同一幅图中的其他区域，再经过模糊、边缘处理等操作便能达到以假乱真的效果。在过去的研究中，同图复制检测技术主要分两类，分别为特征点匹配和块匹配[1]。特征点匹配中较为著名的算法有 SIFT(scale invariant feature transform)[2]、SURF(speed up robust features)[3]和 MIFT(mirror reflection invariant feature transform)[4]等。在块匹配算法中，大部分匹配算法的步骤为将图像分成重叠块或非重叠块，提取每个块的特征进行特征匹配，找出窜改区域。特征提取步骤中，根据目的不同，所选择的特征也不尽相同。许多研究者选择 DCT(discrete cosine transformation)系数[5~7]作为特征以提高准确率，并增加对噪声和压缩的鲁棒性，也有选择基于傅里叶变换[8]或小波变换[9]的频率域特征,可有效降低时间复杂度；有些研究则采用空间域特征，如基于颜色不变性[10]、基于纹理特征[11]、基于灰度特征[12,13]等,相比于频率域特征，空间域特征大幅度降低特征维度，减少计算量，增加效率。这些传统的特征提取方法往往会受图像自身变化影响，提取的特征并不能完全反映出物

基金项目：公安部技术研究计划项目（2014JSYJB007）

作者简介：王梦思（1991-），女，吉林长春人，硕士，主要研究方向为网络安全、模式识别（1041146819@qq.com）；霍宏涛（1972-），男，教授，博士，主要研究方向为网络安全、模式识别；罗霄阳（1992-），男，硕士，主要研究方向为网络安全、模式识别.

体的本质属性，无法获得图像更深层次的信息。

深度学习是近年来研究的热点，相较于传统的模式识别方法，它可以从数据中自动学习特征，试图找到数据的内部结构，发现变量之间的真正关系形式，其特征表达能力要优于传统方法并且对大数据的丰富内在信息更有代表性。深度学习应用范围广泛，在图像处理方面，绝大部分研究集中于图像识别[14]、图像分类[15]或检测图像拼接[16]等方面，并多与分类算法相结合。在图像窜改检测方面，特别是同图复制检测技术还少有应用。

聚类算法在同图复制窜改检测中已有应用，其中应用较多的聚类算法为 K-means[17]，聚类的优势在于降低时间复杂度，减少计算量。应用深度学习特征，结合其他特征和聚类算法，检测并能准确定位复制粘贴窜改区域是本文研究的主要内容。

为提高检测准确率，降低时间复杂度，本文将图像分为若干重叠块，并选择基于稀疏自编码(sparse autoencoder SAE)神经网络提取分块的隐性特征，应用 K-means 聚类算法去除背景定位复制区域，再提取剩余分块纹理特征，使用 K-means 作二次聚类，得到窜改区域检测图。若仍有非窜改区域的异常块，则使用欧氏距离判定和 RANSAC 算法去除异常，从而实现同图复制窜改检测。实验结果表明，该算法在时间效率和准确率上都要优于其他算法，并对 JPEG 压缩也有很好的鲁棒性。

1　稀疏自编码器

自编码神经网络（autoencoder）是一种无监督的学习算法，它使用反向传播算法，让样本的输出值等于输入值，如果输入数据间隐含有某些有联系的特定结构，那么自编码算法就能发现这些有联系数据间的相关性，并在输出层重构出输入数据[15]。图 1 是一个自编码神经网络的简单示意图。其中 $\{x^{(1)},x^{(2)},L\ ,x^{(6)}\}$ 为训练样本集合， $X^{(i)}\in \mathrm{R}$，即输入层 L1，L2 为隐藏层，输出层 L3: $h_{w,b}(x)\approx x$， +1 为偏置项（截距项）系数, W 是权重参数，b 为截距。

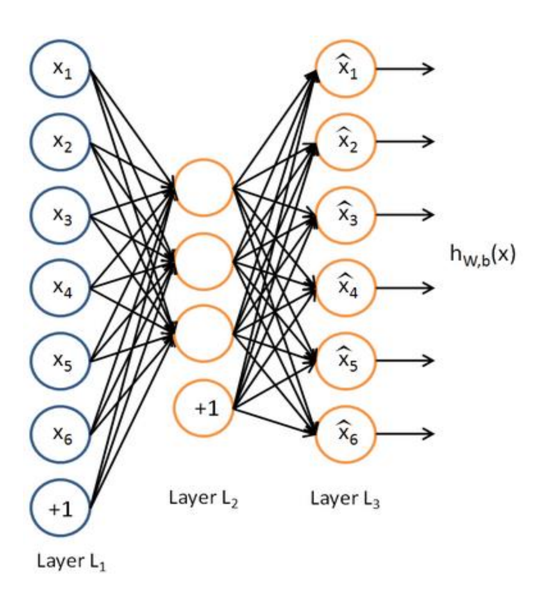


图 1　自编码神经网络示例

在该神经网络中加入一些条件，如限制隐藏层 L2 中神经元的个数，则会出现某种特定结构，该结构可让 L2 重构 L3。设神经元的激活函数是 sigmoid 函数 $\mathrm{f}(x)=\frac{1}{1+e^{-x}}$，当神经元的输出接近于 1 的时候则为激活状态，接近 0 的时候为抑制状态，稀疏性限制使输出接近于 0。若共 $\{(x^{(1)},y^{(1)})\ldots(x^{(m)},y^{(m)}\}$ m 个样例，则公式如下：

$$\hat{\rho}_j=\frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}\left(x^i\right)\right] \tag{1}$$

$a_j^{(2)}$ 表示当输入为 x 情况下隐藏神经元 j 的激活度，$\hat{\rho}_j$ 表示 j 的平均活跃度，加入限制条件，令 $\hat{\rho}_j=\rho$， ρ 是稀疏性参数，其值接近于 0，当隐藏层神经元平均活跃度接近于 ρ 时，平均活跃度则接近于 0，同时加入惩罚因子降低平均活跃度。公式如下：

$$\sum_{j=1}^{s_2}KL(\rho\,\|\,\hat{\rho}_j)=\sum_{j=1}^{s_2}\rho log\frac{\rho}{\hat{\rho}_j}+(1-\rho)log\frac{1-\rho}{1-\hat{\rho}_j} \tag{2}$$

s_2 是隐藏层中隐藏神经元的数量，索引 j 依次代表隐藏层中的每一个神经元，$KL(\rho\,\|\,\hat{\rho}_j)$ 是一个以 ρ 为均值和一个以 $\hat{\rho}_j$ 为均值的两个伯努利随机变量之间的相对熵。当 $\hat{\rho}_j=\rho$ 时 $\mathrm{KL}(\rho\,\|\,\hat{\rho}_j)=0$，这样，总体代价函数为

$$\begin{aligned}J_{sparse}(W,b)=&\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left\|y^i-h_{w,b}\left(x^i\right)\right\|^2+\\&\frac{\lambda}{2}\sum_{l=1}^{n_l}\sum_{q=1}^{s_l}\sum_{p=1}^{s_{l+1}}\left(w_{pq}^{(l)}\right)^2+\beta\sum_{j=1}^{s_2}KL(\rho\,\|\,\hat{\rho}_j)\end{aligned} \tag{3}$$

其中：n_l 为自编码神经网络层数； λ 为规则化系数； β 是控制稀疏性限制惩罚项的系数； $h_{w,b}(x^i)$ 是第 i 组神经网络输出层的输出值。式（3）采用迭代法算取平均激活度 $\hat{\rho}$，当迭代达到设定值或算法收敛时停止。

2　特征聚类算法

2.1　本文算法流程

将 SAE 特征结合聚类算法检测同图复制窜改是本文的主要思想，具体流程如图 2 所示。

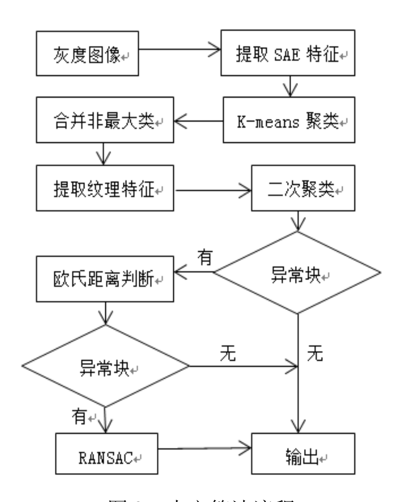


图 2　本文算法流程

总结步骤如下：

a)将图像转为灰度图像后，进行间隔为 d 的重叠分块，提取每块的 SAE 特征；

b)应用 K-means 聚类将平滑区域去除，即排除含最大块数的类，合并剩余类；

c)将剩余类里的块提取纹理特征进行 K-means 二次聚类，根据图像的复杂程度不同，此时有些图像已可以成功检测，检测结果在二次聚类后的某一类中；

d)若有些图像仍存在少量异常块（非窜改区域），使用欧氏距离判断和 RANSAC 算法将异常块去除，保留完整的复制粘贴区域。

2.2 SAE 特征提取

将待检测图像转换为灰度图像后，将图像分成 $n\times n$ 大小的块。为提高效率，采用间隔 d 行 d 列分块。此时，应用稀疏自编码器，输入特征值 n^2 维，输出隐藏层特征矩阵 m^2 维，m 小于 n。

通过在稀疏自编码器中训练大量样本，增加迭代次数,发现样本中的规律，得到最终收敛的隐藏层权重参数矩阵 $\boldsymbol{W}$和截距 $\boldsymbol{b}$。

若设定输入为 16x16=256 维特征的样本，隐藏层输出为 25 维特征，示例如下：

$$\underset{\text{输入层}}{X} \xrightarrow{W^1\ b^1} \underset{\text{输出层}}{Z}$$

$\boldsymbol{X}=[\ ]_{256\times n}$，$\boldsymbol{X}$ 是 $256\times n$ 的矩阵，这里的 n 是指训练样本集中每张图片的所有分块，即将图片分块，块大小为 16×16，归成列向量 256 维，所有图片分块的特征向量组成矩阵 $\boldsymbol{X}$。

$\boldsymbol{W}^1=[\ ]_{25\times256}$，25 表示隐藏层节点数，第 i 行代表每个输入特征与隐藏层第 i 个节点的系数。256 代表输入节点数，第 j 列表示输入特征和第 j 个节点的系数。$\boldsymbol{b}^1=[\ ]_{25\times1}$，$\boldsymbol{Z}=\boldsymbol{W}^1\times\boldsymbol{X}+\boldsymbol{b}^1=[\ ]_{25\times n}$，得到的 $\boldsymbol{Z}$ 为 256 维原始输入向量的 25 维特征表示矩阵，本文将 $\boldsymbol{W^1}$ 和 $\boldsymbol{b^1}$ 统一表示为 $\boldsymbol{W}$。例如 $\boldsymbol{Z_1}$ 为一个块的 25 维特征矩阵(保留两位小数)。

$$Z_1=\begin{bmatrix} 3.18 & 0.60 & 3.32 & -0.59 & 8.29 \\ -2.48 & -11.22 & -7.24 & -3.23 & 1.61 \\ -1.36 & 1.33 & -5.25 & -1.56 & -1.01 \\ 1.29 & -1.43 & 3.85 & -2.33 & -4.88 \\ 0.23 & -6.48 & -3.79 & -2.42 & 2.25 \end{bmatrix}$$

稀疏自编码提取图像的特征为隐藏层单元对图像不同方向和不同位置的边缘检测，相较于一般边缘检测算子如 Laplacian、Prewitt 等，其优势也很明显，稀疏自编码强调图像深层内部特征，着重于图形整体，所以针对外部条件变换具有更好的稳定性，这种特征适用于结合其他算法进行进一步对于图像的操作。

2.3 k-means 聚类算法

k-means 算法将基于特征的对象中任意选择 k 个对象作为初始聚类中心，将剩余对象根据聚类中心的相似度划分 k 隔簇，通过迭代计算新的聚类中心，直到目标函数达到收敛即可。准则函数公式如下，k 为指定的聚类数目，D 为样本数。

$$d=min\sum_{i=1}^{K}\sum_{x\in D_i}dist(D_i,x)^2 \tag{4}$$

目标函数达到簇不发生变化或者设定最大的迭代次数即可终止。算法流程如图 3 所示。

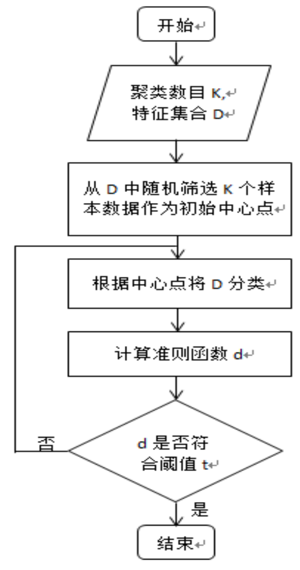


图 3 K-means 算法流程

从图像伪造数据库 CASIA TIDE V 2.0[1]选取若干图像进行测试，将测试图像的稀疏自编码特征进行 k-means 聚类，设定 k=5，根据 k-means 特性，每个聚类应集中分布于图像不同区域，而将含有最大块数的聚类在图像中可视化后发现其分块分布在整幅图像，故将其排除，剩下的 4 个聚类的块合并，得到合并类矩阵 A，可视化的效果如图 4 所示。

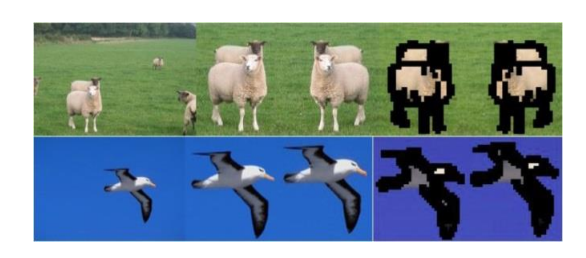


图 4 左：原图；中：窜改图；右：检测结果

从图 4 中可以看出，稀疏自编码特征可完整描述图像边缘信息，若图像只有窜改区域，则直接得出检验结果，若图像中除窜改区域外还包括其他图形，则测试结果如图 5 所示。

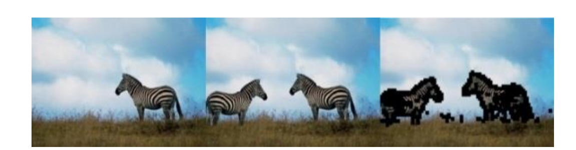


图 5 左：原图；中：窜改图；右：检测结果

1 CASIA Image Tampering Detection Evaluation Database,Ver.2.0,2010,http://forensics.idealtest.org.

从图 5（右）显示结果来看，若图像中含其他图形，合并类里会存在异常块（在非窜改区域）。由此可见，SAE 特征结合 k-means 算法可将大面积的图像平滑区域去除，这样大大减少计算量，即使出现异常块，仅需处理合并类，将图像的异常块去除。

提取合并类 A 里每个分块的基于灰度共生矩阵的纹理特征组成特征矩阵 ***B***，将 ***B*** 进行 K-means 二次聚类。这样做的目的是，根据分块纹理特征的差异，K-means 可较好地将复制粘贴区域块聚类，进一步区分窜改块和异常块。在二次聚类后，会出现两种情况，一是某一类中的结果为无异常块的窜改区域，即完整检测出复制粘贴区域；二是依旧存在少量异常块，需进一步去除。

2.4 去除异常块

根据式（5）算出聚类中两两块特征向量的欧氏距离，判断是否符合设定的阈值，若小于阈值则保留，否则认为两块不匹配。

$$d_{12}=\sqrt{\sum_{i=1}^{n}\left(x_{1i}-x_{2i}\right)^2} \tag{5}$$

将余下匹配块结果点对的坐标为输入，应用 RANSAC 算法去除点对集合中的异常块，保留正常块，进而循环建立原始区域和复制伪造区域之间的仿射变换矩阵，RANSAC 生成的仿射变换矩阵很好地反映了原始区域和窜改区域之间的位置变换关系[11]。

3 实验结果

3.1 针对不同情况准确率的检测

在前期工作中，用稀疏自编码器训练图库里的 70 张复制窜改图像，设定 n=16，m=5，d=4, λ=0.0001，β=3，ρ=0.01，最大迭代数 N=8000,得到的权重参数矩阵 ***W*** 与图像分块矩阵 ***K*** 相乘，***K*** 的一列是一个块的特征向量。本文选取图库里的 40 张同图复制窜改图像进行检测，完整地检测出 37 张图，识别达到 92.5%。

由于 RANSAC 算法的输入量越大，用时越多，而输入量小则会降低准确率，综合考虑算法时间效率和准确率两方面因素，设定输入量即输入块的个数范围 $m\in[300,1000]$，欧氏距离阈值设定范围为 $d\in(0.02,0.10)$。实验结果表明，当 $d\le 0.02$ 时，约 74%的测试图像 $m\le 300$，从而增大漏判数；当 $d\ge 0.10$ 时，约 81%的测试图像 $m\ge 1000$，从而降低时间效率。本文实验阈值初始设定值为 $d=0.05$，根据 m 值的范围，自动调整 d 的大小，调整值为每次 $d=d\pm 0.015$。将检测图像进行图像形态学操作优化处理，最终得到仿真结果如图 6 所示。

同时将图 6 四组图进行 3 种不同程度的 JPEG 压缩。分别设置 JPEG 压缩质量因子(quality factor QF) QF=95,75,50 进行对于压缩图片的检测，将模拟文献 [13]算法的实验结果作为参照。通过计算准确率(precision)和召回率(recall) 来测评。准确率和召回率公式如下：

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \qquad \text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

其中：TP 是指被正确地划分为正例的个数；FP 是指被错误地划分为正例的个数；FN 是指将正例漏判的个数[11]。

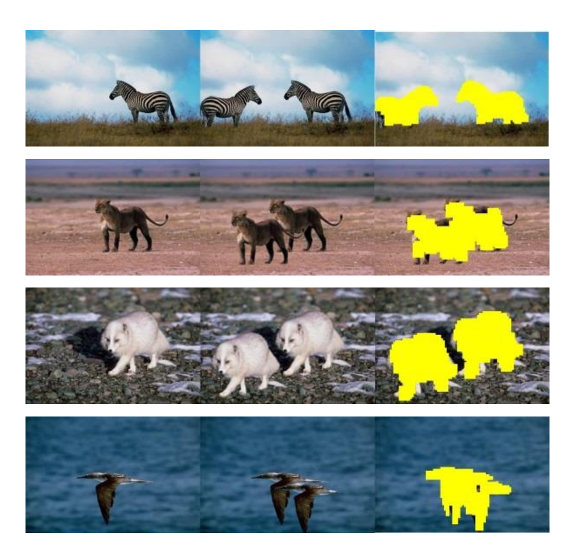


图 6 左：原图；中：窜改图；右：检测图

表 1 为图 6 不同 QF 下的两种算法 precision 值，表 2 为图 6 不同 QF 下的两种算法 recall 值。

表 1 图 6 不同 QF 下的两种算法 precision 值

	QF=100	QF=95	QF=75	QF=50
[13]	0.874	0.816	0.723	0.577
本文	0.930	0.894	0.831	0.791

表 2 图 6 不同 QF 下的两种算法 recall 值

	QF=100	QF=95	QF=75	QF=50
[13]	0.881	0.825	0.753	0.607
本文	0.943	0.906	0.869	0.802

在检测的 40 张图中统计其中 20 张图在 QF=95，90，85，80，75 的准确率和召回率，以确定本文算法针对 JPEG 压缩的鲁棒性，如图 7 所示。

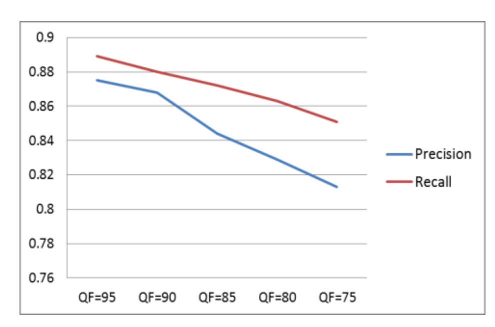


图 7 不同 QF 下的本文算法 precision 和 recall 值

可见针对 JPEG 压缩，SAE 特征的稳定性较好，算法检测率也优于其他算法。

3.2 时间效率检测

除了检测准确率提升外，本文算法的突出优势在于极大地降低时间复杂度，提升计算效率。本文算法的平均用时计算分三种情况：

a)图像背景为平滑区域，除背景外只有复制粘贴窜改区域，

此情况仅需提取 SAE 特征并应用 K-means 算法便可完整地检测出，其平均用时 1.87 s。

b)图像中除窜改区域，还有其他图形，此情况除了用情况 a)的步骤外，还需提取纹理特征，使用 K-means 二次聚类，此时某一聚类中有完整的窜改检测结果，其平均用时 18.26 s。

c)二次聚类后，仍有异常块，此时需使用欧氏距离和 RANSAC 算法去除异常块，其平均用时如表 3 所示，并与文献[13]做对比。

表 3　两种算法的平均用时/s

	文献[13]	本文	提升率
特征提取与匹配	109.55	19.37	82%
去除异常块	416.37	127.41	69%
总共	525.92	146.78	72%

4　结束语

为了提升块匹配算法的准确率和时间效率，本文提出将深度学习特征和 K-means 聚类算法相结合，从而检测同图复制窜改区域。实验结果表明，在准确率上，不同压缩程度的图像检测结果皆优于其他算法，而时间效率也大幅度提升，这表明深度学习在同图复制窜改检测中具有很好的应用性。在未来的工作中，可进一步探索自编码神经网络在此领域的应用，如图像窜改区域经缩放或旋转等处理。

参考文献：

[1] Christlein V, Riess C, Jordan J, et al. An evaluation of popular copy-move forgery detection approaches [J]. IEEE Trans Information Forensics & Security, 2017 (6): 1841-1854.

[2] Pan Xunyu, Lyu S W. Region duplication detection using image feature matching [J]. IEEE Trans on Information Forensics & Security, 2010, 5 (4): 857-867.

[3] Bay H, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110 (3): 346-359.

[4] Jaberi M, Bebis G, HussainM, et al. Improving the detection and localization of duplicated regions in copy-move image forgery [C]// Proc of the 18th IEEE International Conference on Digital Signal Processing. 2013: 2013: 1-6.

[5] Kang X, Wei S. Identifying tampered regions using singular value decomposition in digital image forensics [C]// Proc of International Conference on Computer Science & Software Engineering. 2008: 926-930.

[6] Kumar S, Desai J, Mukherjee S. A fast DCT based method for copy move forgery detection [C]// Proc of the 2nd IEEE International Conference on Image Information Processing. 2014: 649-654.

[7] Wandji N D, Sun X, Kue M F. Detection of copy-move forgery in digital images based on DCT [J]. Computer Science, 2013, 111 (1): 148-65.

[8] Ketenci S, Ulutas G. Copy-move forgery detection in images via 2D-Fourier transform [C]// Proc of International Conference on Telecommunications & Signal Processing. 2013: 813-816.

[9] Khan E S, Kulkarni E A. An efficient method for detection of copy-move forgery using discrete wavelet transform [J]. International Journal on Computer Science & Engineering, 2010, 2 (5): 1801-1806.

[10] Li Jing, Chao Shao. Image copy-move forgery detecting based on local invariant feature [J]. Journal of Multimedia, 2012, 7 (1) .

[11] 王任华，霍宏涛，蒋敏. RANSAC 算法在同图复制鉴定中的应用研究 [J]. 计算机应用研究, 2014, 31 (7): 2209-2212.

[12] Lynch G, Shih F Y, Liao H M. An efficient expanding block algorithm for image copy-move forgery detection [J]. Information Sciences, 2013, 239 (4): 253-265.

[13] Chen C C, Wang H, Lin C S. An efficiency enhanced cluster expanding block algorithm for copy-move forgery detection [C]// Proc of International Conference on Intelligent Networking and Collaborative Systems. 2015.

[14] Wu M, Chen L. Image recognition based on deep learning [C]// Proc of Chinese Automation Congress. 2016.

[15] 王勇，赵俭辉，章登义，等. 基于稀疏自编码深度神经网络的林火图像分类 [J]. 计算机工程与应用, 2014, 50 (24): 173-177.

[16] Rao Y, Ni J. A deep learning approach to detection of splicing and copy-move forgeries in images [C]// Proc of IEEE International Workshop on Information Forensics & Security. 2017: 1-6.

[17] Fadl S M, Semary N A. A proposed accelerated image copy-move forgery detection [C]// Proc of Visual Communications & Image Processing Conference. 2014: 253-257.